Boston University

OpenBU **http://open.bu.edu**

Biomedical Engineering ENG: Biomedical Engineering: Scholarly Papers

2009-1

TB Database: An Integrated Platform for Tuberculosis Research

Reddy, T. B. K., Robert Riley, Farrell Wymore, Phillip Montgomery, Dave DeCaprio, Reinhard Engels, Marcel Gellesch, Jeremy Hubble, Dennis Jen, Heng Jin, Michael Koehrsen, Lisa Larson, Maria Mao, Michael Nitzberg, Peter Sisk, Christian Stolte, Brian Weiner, Jared White, Zachariah K. Zachariah, Gavin Sherlock, James E. Galagan, Catherine A. Ball, Gary K. Schoolnik. "TB database: an integrated platform for tuberculosis research" Nucleic Acids Research 37(Database issue): D499-D508. (2009)
https://hdl.handle.net/2144/3030
Boston University

Published online 3 October 2008 Nucleic Acids Research, 2009, Vol. 37, Database issue D499–D508
doi:10.1093/nar/gkn652

TB database: an integrated platform for tuberculosis research

T. B. K. Reddy[1,*], Robert Riley[2], Farrell Wymore[1], Phillip Montgomery[2], Dave DeCaprio[2], Reinhard Engels[2], Marcel Gellesch[2], Jeremy Hubble[3], Dennis Jen[2], Heng Jin[1], Michael Koehrsen[2], Lisa Larson[2], Maria Mao[3], Michael Nitzberg[1], Peter Sisk[2], Christian Stolte[2], Brian Weiner[2], Jared White[2], Zachariah K. Zachariah[1], Gavin Sherlock[3], James E. Galagan[2,4,5], Catherine A. Ball[1] and Gary K. Schoolnik[6]

[1]Department of Biochemistry, Stanford University, CA 94305, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, [3]Department of Genetics, Stanford University, CA 94305, [4]Department of Biomedical Engineering, Boston University, Boston, MA 02215, [5]National Emerging Infectious Diseases Lab, Boston University, Boston MA 02118 and [6]Department of Microbiology & Immunology, Stanford University, CA 94305, USA

Received August 14, 2008; Revised September 17, 2008; Accepted September 18, 2008

ABSTRACT

The effective control of tuberculosis (TB) has been thwarted by the need for prolonged, complex and potentially toxic drug regimens, by reliance on an inefficient vaccine and by the absence of biomarkers of clinical status. The promise of the genomics era for TB control is substantial, but has been hindered by the lack of a central repository that collects and integrates genomic and experimental data about this organism in a way that can be readily accessed and analyzed. The Tuberculosis Database (TBDB) is an integrated database providing access to TB genomic data and resources, relevant to the discovery and development of TB drugs, vaccines and biomarkers. The current release of TBDB houses genome sequence data and annotations for 28 different *Mycobacterium tuberculosis* strains and related bacteria. TBDB stores pre- and post-publication gene-expression data from *M. tuberculosis* and its close relatives. TBDB currently hosts data for nearly 1500 public tuberculosis microarrays and 260 arrays for *Streptomyces*. In addition, TBDB provides access to a suite of comparative genomics and microarray analysis software. By bringing together *M. tuberculosis* genome annotation and gene-expression data with a suite of analysis tools, TBDB (http://www.tbdb.org/) provides a unique discovery platform for TB research.

INTRODUCTION

In humans, tuberculosis (TB) is caused by the bacterium *Mycobacterium tuberculosis* and primarily targets the lungs (as pulmonary TB), but can also affect other organs, including the brain and meninges, lymph nodes, bone and joints, the genitourinary system and the intestine and liver. TB is today the second highest cause of death from infectious diseases after HIV/AIDS (1) and is the biggest killer of people infected with HIV (2). The World Health Organization's most recent global data (from 2005) show that every year 8 million people become ill with tuberculosis and 2 million people die of the disease. A third of the world's population has been exposed to TB, making this disease one of the greatest global health challenges facing us today (3). A remarkable feature of TB is its ability to enter an asymptomatic latent phase lasting years or even decades. Activation of a latent infection can be precipitated by changes in the physiological and immune status of the host owing to declining cell-mediated immunity associated with senescence, malnutrition and diabetes or the occurrence of other diseases, especially HIV/AIDS (4). Chemotherapy for active TB due to drug-sensitive strains entails the use of multiple antibiotics administered for 6 months. This complicated and frequently toxic treatment regimen often results in poor patient compliance. This in turn has led to the emergence of antibiotic resistant strains that require longer treatment courses, the use of less effective and more toxic drugs and higher failure rates (5). As a result, TB remains a widespread and deadly disease whose control will require more

*To whom correspondence should be addressed. Tel: 650 736 0075; Fax: 650 724 3701; Email: tbreddy@stanford.edu

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

effective public health measures and the development of new drugs and vaccines. Recent developments in genomics and the availability of the complete *M. tuberculosis* genome sequence (6) has led to the use of genome-wide expression profiling and comparative genomics methods to better understand *M. tuberculosis* pathology, latency, emerging drug resistance and evolution. However, despite the wide-spread use of functional and comparative genomics to study *M. tuberculosis*, there has been no single repository for these large-scale datasets, complete with high-quality experimental annotation, and connected to up-to-date gene annotation and comparative genomic information. Instead, much of these data have been located in disparate sites like GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes (7) and MGDD: *M. tuberculosis* genome divergence database (8) that employ diverse and often incompatible formats and analytical tools. The Tuberculosis Database (TBDB) was developed to address this gap. TBDB uses software from the Stanford Microarray Database (SMD) (9) and the Broad Institute's Calhoun system (10,11), and houses gene-expression data paired with genome sequence and annotation data. Uniting experimental data with genome sequence data enables researchers to ask complex questions and draw inferences that would otherwise be impossible by looking at individual small datasets. In this context, TBDB brings together powerful genomics tools to advance *M. tuberculosis* research in ways that will contribute to the identification of new drug targets, vaccine antigens, diagnostics and host biomarkers.

TBDB OVERVIEW

TBDB is an integrated database that houses both annotated genome sequence data and microarray and RT–PCR expression data from *in vitro* experiments and TB-infected tissues. TBDB houses genome sequence data for several *M. tuberculosis* strains as well as data for numerous related species. These data and annotations include publicly available sequences from a number of sequencing centers and groups, including sequences being produced by the Broad Institute's Microbial Sequencing Center. The microarray data within TBDB are predominantly from *M. tuberculosis*, but we are in the process of incorporating *in vivo* data from infected host tissues (principally human, primate and murine) into TBDB. Experimental data may be deposited into TBDB by any TB researcher prior to publication providing prepublication access to tools for the analysis, annotation, visualization and sharing of data. The data are then made public at the author's request or following publication, whichever is first. In addition, TBDB curators search the literature for publications containing relevant TB or host microarray data. The primary data are then requested from the authors of such publications and are entered into TBDB, where the experiments are annotated and made public so other researchers can reanalyze the data (often in conjunction with other datasets within TBDB) using TBDB tools. Table 1 lists TBDB statistics, including the number of annotated genomes in TBDB, microarray experiments, publications and other data types.

Table 1. Summary of TBDB data content (as of September 2008)

TBDB data statistics	
Number of genomes	28
Number of all microarrays	~5500
Number of public microarrays	~1800
Number of publications	27
Number of experiment sets	160

The first route of entry into TBDB is the *Quick Search* feature, which allows a user to search all objects in TBDB by gene name, gene sequence name, author name, title or any other keyword. The result page of a *Quick Search* provides a count of genes, microarray experiments, operons, gene families and other database objects that match the query. Links from this results page provide direct access to pages with detailed information about particular objects, such as the Gene Detail and Publication pages. *Quick Search* is available at the top of every TBDB page, and thus provides an easily accessible single integrated access point to all genome annotation and expression data in TBDB.

TBDB GENOMES

TBDB currently houses genome sequence data for *M. tuberculosis* strain H37Rv (a standard prototype strain long used for experimental and animal infection studies), as well as other *M. tuberculosis* strains and bacteria from related taxa, focusing on members of the Actinomycetes family of high G+C content, Gram-positive organisms of which *M. tuberculosis* is a member. These genomes sequences have been annotated with a variety of genomic features including genes, operons, sequence similarity to GenBank sequences using BLAST (12), transfer RNAs using tRNAScan (13), protein domains and families using PFAM (14) and noncoding RNAs based on RFAM (15). Known immune epitopes have also been mapped through collaboration with BioHealthBase (16). A suite of analytical tools is also provided to allow comparative genomic analysis of *M. tuberculosis*. Table 2 lists the genomes in TBDB for which sequence data are available along with their size and the number of annotated genes. Access to the annotated genome sequences and comparative data is provided through several search interfaces, some of which are described subsequently.

Feature detail pages

All information about annotated features on any genome sequence is available through Feature Detail pages, of which the Gene Detail page is the most common example (Figure 1). Information presented in the Gene Detail page is organized into different sections. These include, *Gene Info, Gene Expression, Functional Annotation, Transcript Info, Sequence* and genome display options. The Gene Info section provides complete details about *Locus Name, Gene Symbol, Synonyms, Gene Name, Gene*

Table 2. List of annotated genomes in TBDB

Organism	Size (mb)	Genes
M. tuberculosis H37Rv	4.41	3999
M. tuberculosis CDC1551	4.4	4189
M. tb. F11 (finished)	4.42	3959
M. tb. C	4.38	3851
M. tb. Haarlem	4.4	3866
M. bovis AF2122/97	4.35	3920
M. bovis BCG	4.37	3952
M. leprae TN	3.27	1605
M. avium 104	5.48	5120
M. avium k10	4.83	4350
M. smegmatis MC2 155	6.99	6716
M. marinum	6.64	5423
M. ulcerans Agy99	5.63	4160
M. vanbaalenii PYR-1	6.49	5979
M. sp. KMS	6.26	5975
M. sp. MCS	5.71	5391
Rhodococcus sp. RHA1	9.7	9145
Nocardia farcinica IFM 10152	6.02	5683
Corynebacterium glutamicum ATCC 13032	3.28	3057
C. diphtheriae NCTC 13129	2.49	2272
C. efficiens YS-314	3.15	2950
C. jeikeium K411	2.48	2120
Streptomyces avermitilis MA-4680	9.12	7673
S. coelicolor A3(2)	8.67	7825
Propionibacterium acnes KPA171202	2.56	2297
Acidothermus cellulolyticus 11B	2.44	2157
Bifidobacterium longum NCC2705	2.26	1727
Rhodobacter sphaeroides	4.6	4242

Product Names, Gene Family, Location, Protein Domains, External Links to related databases including TubercuList (17), TB Structural Genomics Consortium (TBSGC) Protein Structure Information (18) and the Proteome 2D-PAGE Database. Figure 1 shows the gene detail page for *dosR (devR*, Rv3133c), which encodes the response regulator of a two-component signal transduction system that tightly controls a well-studied *M. tuberculosis* regulon that is activated by oxygen limitation or exposure to nitric oxide (19).

Genome visualization and comparative analysis

Researchers can retrieve DNA or protein sequence for segments of any of the genome sequences in TBDB from many locations within the site, including the *Browse Regions* search tool. The sequences can then be visualized using a number of different tools. The *Argo Genome Browser* (an interactive applet) and the *Feature* Map (a lighter weight version of the *Argo Genome Browser*) provide linear views of genome sequences along with all associated annotated features. Argo in particular provides a dynamic interface to visualizing genome data that allows users to zoom from whole chromosomes to individual nucleotides, navigate within sequences, and select individual features to retrieve additional information. A *Circular Genome Viewer* provides a circular plot of genome sequences along with a plot of the density of particular features, GC content and GC skew. Finally, the *Genome Map* tool provides a dynamic linear view of one or more genome sequences and associated annotations, and displays conserved synteny between the displayed genomes for regions selected by the user (Figure 2).

An additional number of tools are also provided specifically for comparative analyses between genome sequences, including the *Synteny Map*, *Dot Plot*, *Operon Browser* (Figure 3) and *Gene Family Search*. The *Synteny Map* uses precomputed genome alignments to graphically display regions of genomic similarity between a single reference genome and one or more other genomes—in effect providing the results of an *in silico* genome hybridization between sets of genomes. Using this tool, the user can select regions of interest and then click a region to zoom in and view genes, genome sequence, and features. The *Dot Plot* displays a navigable map of computed synteny between genomes in the form of dot-plot lines. When comparing multiple genomes, the color of the plotted synteny indicates which genome is aligned to the reference at that position. The *Operon Browser* is a tool that simultaneously displays the expression correlation between genes in a genomic region of the *M. tuberculosis* H37Rv strain while showing syntenic gene order of orthologs in related species. A heatmap derived from expression correlation data is provided along with an alignment of syntenic areas. Mousing over the genes provides additional information such as locus ID, gene symbol and description. Color coding of genes indicate orthologous relationships across different species. Finally, the *Gene Family Search* displays phylogenetic trees and sequence alignments of predicted orthologous gene families within the genome sequences in TBDB. The basic search feature lets the user choose the number of genomes to query and whether to limit the search to strict orthologs or not. In addition, an advanced search option chooses which genomes to include or exclude.

TBDB GENE EXPRESSION DATA

TBDB houses public and prepublication microarray and RT–PCR expression data. Public data are freely accessible and can be downloaded or reanalyzed using TBDB analysis tools. Access to prepublication data is restricted to the researchers who generated the data until they publish or decide to make their data public. TBDB users can establish a free user account to enter microarray data, share prepublication microarray or RT–PCR data with colleagues or store datasets for analysis in a data repository. Data in the repository can be shared with other researchers at the discretion of the TBDB user.

Expression data in TBDB can be accessed by searching for data from individual microarrays or RT–PCR assays or by searching for data from a publication. For a novice user, the publication search is an easy place to start exploring expression data in TBDB. The expression *Basic Search* is an interactive search option that queries TBDB via publication, organism or dataset. The expression *Advanced Search* finds microarray data by experimenter, category, subcategory and organism. The *Gene Search for Expression* searches for genes or reporter sequences used on microarrays. Reporter sequences correspond to a piece of DNA deposited on a microarray slide.

M. tuberculosis H37Rv: Rv3133c two component transcriptional regulatory protein devR (probably luxR/uhpA-family)

<< Previous Gene on Contig (Rv3132c) Next Gene on Contig (Rv3134c) >>

Gene Info Collapse

Locus	Rv3133c
Gene Symbol	devR
Synonyms	Rv3133,dosR
Gene Name	Rv3133c two component transcriptional regulatory protein devR (probably luxR/uhpA-family)
Gene Product Names	Rv3133c two component transcriptional regulatory protein devR (probably luxR/uhpA-family)
Gene Family	two component transcriptional regulator family
Location	M. tuberculosis H37Rv: 3499262-3499915 -
Length	Gene: 654 nt Protein: 217 aa
Protein Domains	0 20 41 62 83 103 124 145 166 186 207 Response regulator receiver domain Bacterial regulatory c Start Stop Bacterial regulatory proteins, luxR family 147 204 Response regulator receiver domain 2 116 Click to find all genes with that domain
External Links	TubercuList TBSGC Protein Structure Information Proteome 2D-PAGE Database - Search Search Google Scholar

Gene Expression Collapse

Expression History	Reporters: RV3133C-pcr RV3133C-oligo T003133_01-oligo

Functional Annotations Collapse

Functional Annotations	devS-Rv3134c Operon

Transcript Info Collapse

Transcripts	3' 5' [Details] *Product name* : Rv3133c two component transcriptional regulatory protein devR (probably luxR/uhpA-family) This transcript has the following issue(s): ► Missing 5 prime End

Tools Collapse

Display Options	► Feature Map ► Argo Applet ► Desktop Argo (requires Java Web Start) ► Genome Sequence of Locus ► Show on Dotplot ► Show in Operon Browser
Tools	► Local BLAST with Protein Sequence ► Local BLAST with Coding Sequence ► NCBI BLAST with Protein Sequence ► NCBI BLAST with Coding Sequence

Overlapping Features Expand

Sequence Expand

Figure 1. TBDB Gene Detail page. The Gene Detail page provides at-a-glance information for a given gene, including known names and synonyms, predicted function(s) and protein domains. It also serves as a jumping off point to various sequence tools, and to expression data for that gene. In addition, it provides several links to external resources such as TubercuList, TBSGC Protein Structure Information, Proteome 2D-PAGE Database at Max Planck Institute.

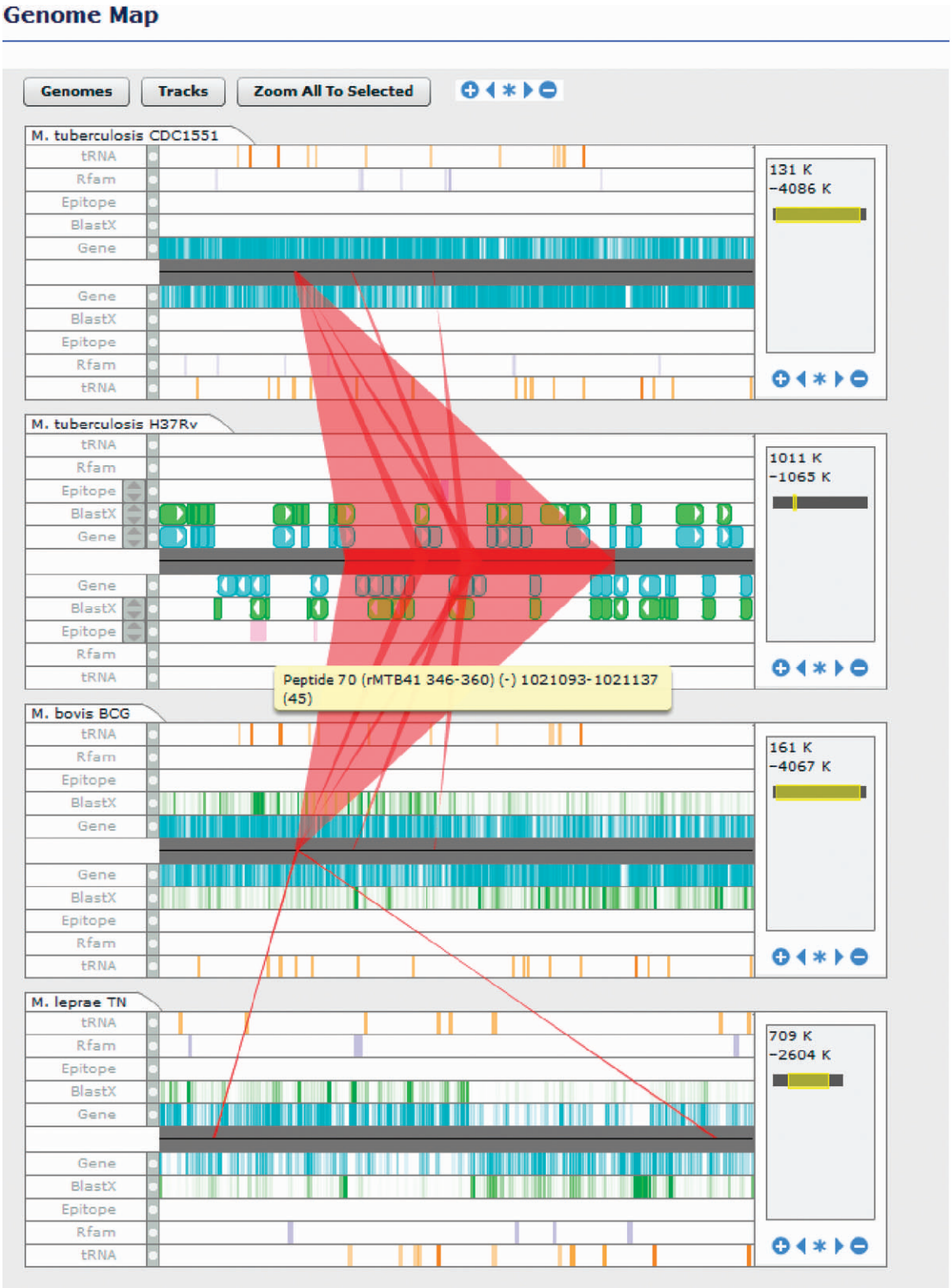

Figure 2. Genome Map tool. This tool provides a linear view of one or more genome sequences and associated annotations as well as conserved synteny between genomes. Annotations are provided as tracks above (forward strand) and below (reverse strand) the midline. When zoomed out, annotations are viewed as density plots; when zoomed in individual features are displayed. Users may select regions of a genome sequence by dragging along the midline. Syntenic regions in the other sequences associated with the selection are then displayed as red bands.

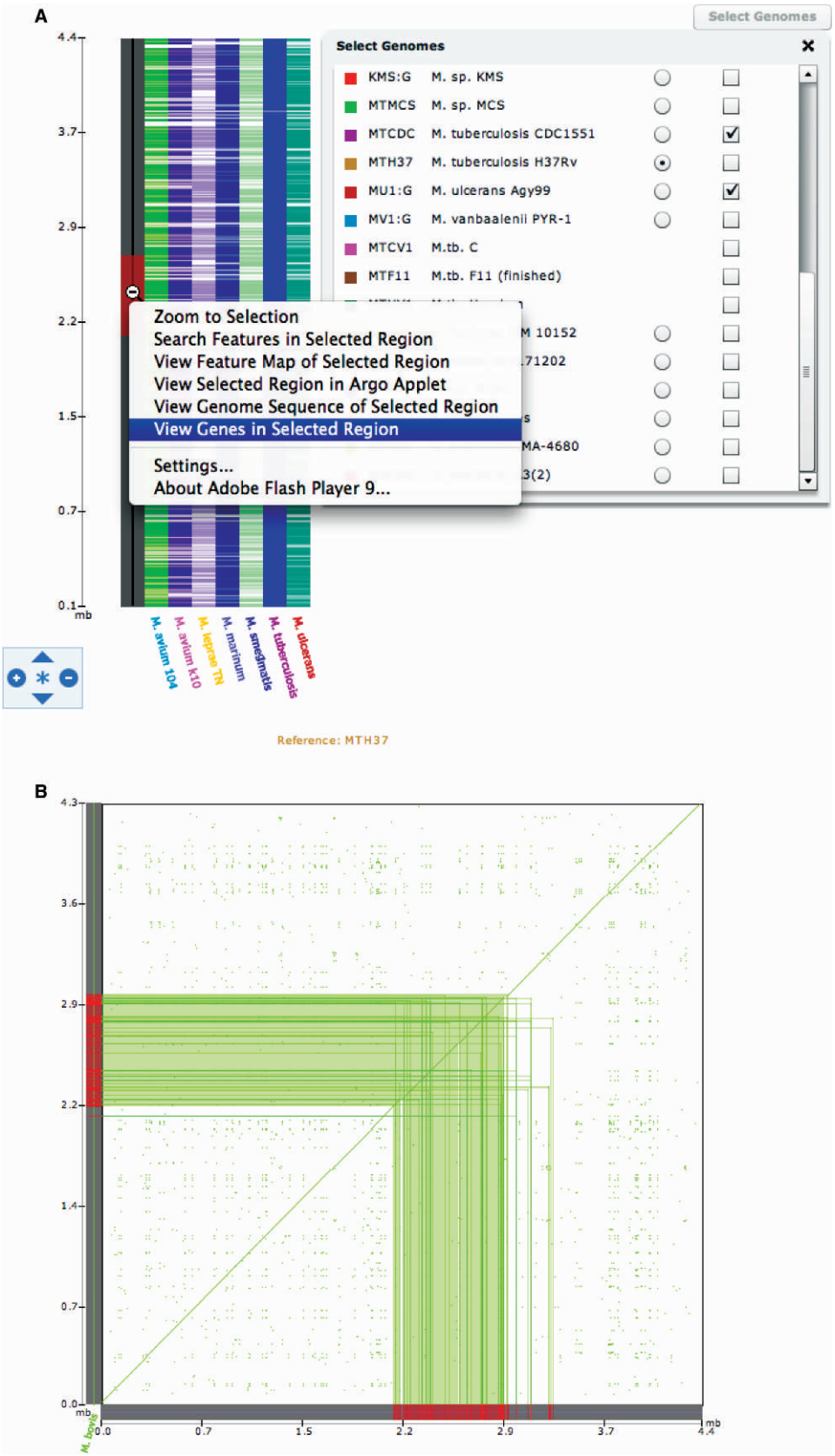

Figure 3. Comparative genome analysis. The *Genomes Synteny Map* (**A**), *Dot Plot* (**B**) and *Operon Map Browser* (**C**) provide different ways to access comparative genomic data between *M. tuberculosis* reference genome and selected related species. These tools provide an interactive means to explore comparative genomic data.

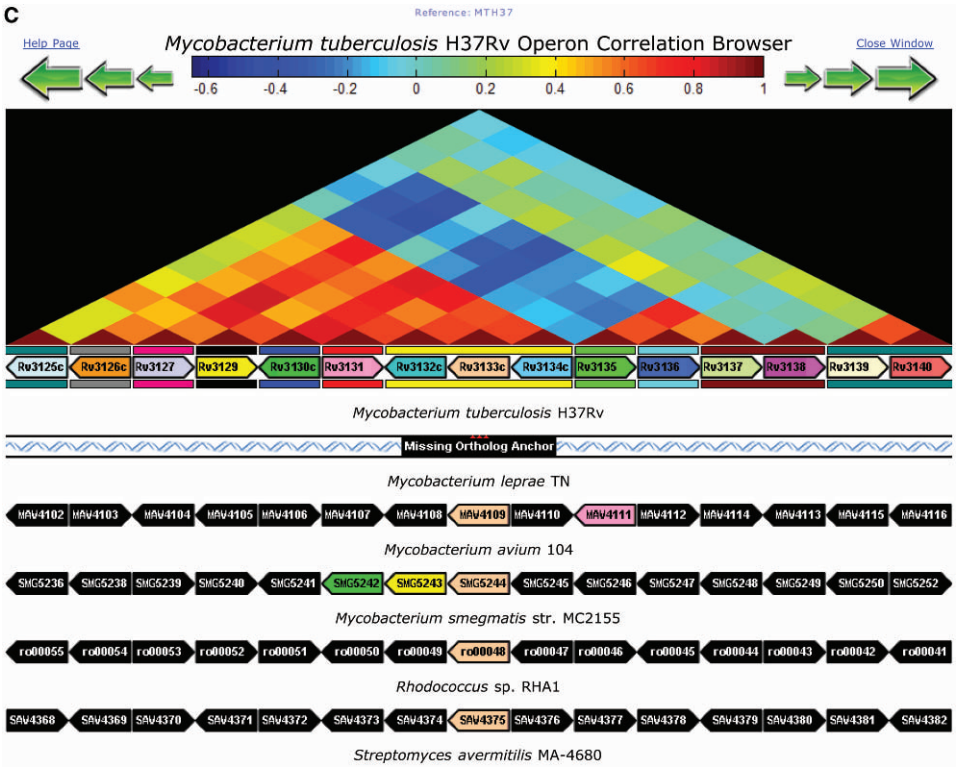

Figure 3. Continued.

This search returns all microarray spots associated with a reporter sequence or gene, and the search results link to the *Spot History* page that lets users explore all associated microarray data.

Expression connection

Using *Expression Connection*, researchers can visualize and explore clustered microarray datasets from publications whose data are present within TBDB. Clustering organizes expression data for genes or reporter sequences into groups that have similar expression profiles. This enables a user to directly view and explore already clustered data within TBDB without needing to go through the data analysis pipeline. As shown in Figure 4, a publication detail page can be accessed by following TB Expression → Gene Expression Publications → 'Data in TBDB'. Interactive clustered data images for a publication can be navigated using GeneXplorer (20), which provides views of the most correlated genes for a gene of interest or searches for genes using text queries (Figure 4). Thus, this option enables a user to explore and interrogate TBDB for expression data from publications.

Data analysis

TBDB provides a suite of microarray data analysis tools for its users. All tools are freely available to analyze both public and prepublication data in TBDB. A typical data analysis process at TBDB involves several steps in the following order: Experiment Selection → Gene Selection and Annotation → Data Filtering Options → Data Retrieval → Gene Filtering → Clustering and Image Generation. At each step, a user is presented with various options that allow them to filter and cluster the data according to their needs. For example, a user can employ either the *Basic* or *Advanced Expression Search* to choose a set of microarray data for further analysis. Clicking on the 'Data Retrieval and Analysis' option invokes the data analysis pipeline, where a user can select various microarray data filtering and transformation options. Many microarray data analysis tools can be applied to datasets, including hierarchical clustering, imputation of missing values, Gene Set Enrichment Analysis (21), Singular Value Decomposition (22) and pathway analyses. All SMD analysis tools [many described previously (9)] have been made available through TBDB. At each step in the data analysis, pipeline a link to a relevant 'Help' page is provided, which explains in detail the various available options. In addition, the TBDB data repository provides access to the suite of gene-expression analysis tools provided through the Gene Pattern software (23).

Literature curation

Curating microarray expression data from publications is an important part of TBDB's efforts. We actively search PubMed for relevant publications containing microarray experiments, then obtain the raw data from researchers and load them into TBDB, with detailed experimental annotations.

Publication Search

Id	Title	Authors	PubMed Link	Full Text Online	Data within database	Web Supplement	Repository
616	The Mycobacterium tuberculosis ECF sigma factor sigmaE: role in global gene expression and survival in macrophages.	Manganelli R , Voskuil MI , Schoolnik GK , Smith I	PubMed	Full Text	Data in TBDB	Web Supplement	GEO
632	The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action.	Boshoff HI, Myers TG, Copp BR, McNeil MR, Wilson MA, Barry CE 3rd	PubMed	Full Text	Data in TBDB	Web Supplement	GEO
633	Differential gene expression between Mycobacterium bovis and Mycobacterium tuberculosis.	Rehren G, Walters S, Fontan P, Smith I, Zarraga AM	PubMed	Full Text	Data in TBDB		GEO
631	Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program.	Voskuil MI, Schnappinger D, Visconti KC, Harrell MI, Dolganov GM, Sherman DR, Schoolnik GK	PubMed	Full Text	Data in TBDB	Web Supplement	GEO

Publication Detail Page

Rustad TR, et al. (2008) PLoS ONE 3(1):e1502

The enduring hypoxic response of Mycobacterium tuberculosis.

Rustad TR, Harrell MI, Liao R, Sherman DR

BACKGROUND: A significant body of evidence accumulated over the last century suggests a link between hypoxic microenvironments within the infected host and the latent phase of tuberculosis. Studies to test this correlation have identified the M. tuberculosis initial hypoxic response, controlled by the two-component response regulator DosR. The initial hypoxic response is completely blocked in a dosR deletion mutant.

METHODOLOGY/PRINCIPAL FINDINGS: We show here that a dosR deletion mutant enters bacteriostasis in response to in vitro hypoxia with only a relatively mild decrease in viability. In the murine infection model, the phenotype of the mutant was indistinguishable from that of the parent strain. These results suggested that additional genes may be essential for entry into and maintenance of bacteriostasis. Detailed microarray analysis of oxygen starved cultures revealed that DosR regulon induction is transient, with induction of nearly half the genes returning to baseline within 24 hours. In addition, a larger, sustained wave of gene expression follows the DosR-mediated initial hypoxic response. This Enduring Hypoxic Response (EHR) consists of 230 genes significantly induced at four and seven days of hypoxia but not at initial time points. These genes include a surprising number of transcriptional regulators that could control the program of bacteriostasis. We found that the EHR is independent of the DosR-mediated initial hypoxic response, as EHR expression is virtually unaltered in the dosR mutant.

CONCLUSIONS/SIGNIFICANCE: Our results suggest a reassessment of the role of DosR and the initial hypoxic response in MTB physiology. Instead of a primary role in survival of hypoxia induced bacteriostasis, DosR may regulate a response that is largely optional in vitro and in mouse infections. Analysis of the EHR should help elucidate the key regulatory factors and enzymatic machinery exploited by M. tuberculosis for long-term bacteriostasis in the face of oxygen deprivation.

Associated Links

Full Text Online | PubMed Link | GEO | Web Supplement | Publication Meta Data

Associated Experiment Sets	Options
H37Rv wild type strain grown under normal vs hypoxic (0.2% Oxygen) conditions	
H37dosR mutant strain grown under normal vs hypoxic (0.2% Oxygen) conditions	

Clustered Data

TB Database
an integrated platform for TB drug discovery

geneXplorer

Put your mouse over elements to see more information here

Search for in All fields go Radar 100% wide change

Zoom

< 1/16 | < 1/ 8 | < 1/ 4 | < 1/ 2 | 1:1 | > 2.0 | > 4.0 | > 8.0 | > 16.0

RV3776		conserved hypothetical protein	conserved hypotheticals (3)
RV3074		conserved hypothetical protein	conserved hypotheticals (3)
RV1334		conserved hypothetical protein	conserved hypotheticals (3)
RV3407		conserved hypothetical protein	conserved hypotheticals (3)
RV3408		conserved hypothetical protein	conserved hypotheticals (3)
RV2391	nirA	probable nitrite reductase/sulphite reductase	intermediary metabolism and respiration (6)
RV1195	PE13	PE-family protein	PE/PPE (3)
RV2124C	metH	5-methyltetrahydrofolate-homocysteine	intermediary metabolism and respiration (3)
RV3519		hypothetical protein	unknown (3)
RV3220C		sensor histidine kinase	regulatory proteins (3)
RV1298	rpmE	50S ribosomal protein L31	information pathways (3)
RV2009		conserved hypothetical protein	conserved hypotheticals (3)
RV3252C	alkB	possible alkane-1 monooxygenase	intermediary metabolism and respiration (3)
RV3251C	rubA	rubredoxin A	intermediary metabolism and respiration (3)
RV3249C		transcriptional regulator (TetR/AcrR family)	regulatory proteins (3)
RV2920C	amt	putative ammonium transporter	cell wall and cell processes (3)
RV3741C		possible monooxygenasemonoxygenase-a	intermediary metabolism and respiration (3)
RV3083		probable monooxygenase	intermediary metabolism and respiration (3)
RV2720	lexA	LexA, SOS repressor protein	regulatory proteins (3)
RV0824C	desA1	acyl-[ACP] desaturase	lipid metabolism (3)
RV1094	desA2	acyl-[ACP] desaturase	lipid metabolism (3)
RV0823C		transcriptional regulator (NifR3/Smm1 family)	regulatory proteins (3)
RV0512	hemB	[delta]-aminolevulinic acid dehydratase	intermediary metabolism and respiration (3)

Figure 4. Publication microarray data and expression connection. Researchers can access the full raw microarray data associated with a publication, either for download, or retrieval through the data retrieval and analysis pipeline. In addition, users can explore clustered microarrays data, whereby they can search for particular genes, or identify which genes show coexpression across a particular dataset.

FUTURE DIRECTIONS

We are working to increase the quality and quantity of data within TBDB and to incorporate additional data types. One of our priorities is to acquire host expression data from *M. tuberculosis*-infected tissues (mouse, primate and human), and we also plan to expand TBDB's capacity to house and analyze RT–PCR data and will develop tools for comparative analysis of RT–PCR and microarray expression data. We will also implement tools such as GO::TermFinder (24), which allows users to determine whether there are biological themes associated with a list of genes of interest, and tools for the analysis of replicate microarray experiments. We are also working to improve the depth and quality of our genome annotations. We are currently curating TB literature and associating these data with genes and other genomics features. Moreover, we have implemented and will deploy a community annotation infrastructure to allow TB researchers to submit additions and improvements to existing annotations through the TBDB website. We are also using the comparative sequence integrated into TBDB to improve on the accuracy of structural gene annotations and to predict additional potential noncoding genes. Finally, as new TB sequences are produced by the Broad Microbial Sequencing Center, they will be deposited and made publicly available in TBDB. Ultimately, we hope that TBDB will serve as a community hub for TB research; a TB research community information page will be implemented with a listing of TB research labs and colleagues; this will also provide a forum for the community of users including feedback and suggestions from the community that will help us better serve them.

CONCLUSION

TBDB contains annotated genome and expression (microarray and RT–PCR) data and a suite of data analysis tools designed to serve as a unique resource for TB research and for the discovery of new drugs, vaccines and biomarkers. Data within the TBDB and all analysis tools are freely available to researchers. Only prepublication gene-expression data require a password.

ACKNOWLEDGEMENTS

We are grateful to the research community for their valuable input and suggestions in building and maintaining this database.

FUNDING

The Bill and Melinda Gates Foundation. Funding for open access charge: The Bill and Melinda Gates Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Arentz,M. and Hawn,T.R. (2007) Tuberculosis infection: insight from immunogenomics. *Drug Discov. Today*, **4**, 231–236.
2. Corbett,E.L., Watt,C.J., Walker,N., Maher,D., Williams,B.G., Raviglione,M.C. and Dye,C. (2003) The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch. Intern. Med.*, **163**, 1009–1021.
3. Young,D.B., Perkins,M.D., Duncan,K. and Barry,C.E. III. (2008) Confronting the scientific obstacles to global control of tuberculosis. *J. Clin. Invest.*, **118**, 1255–1265.
4. Flynn,J.L. and Chan,J. (2001) Tuberculosis: latency and reactivation. *Infect Immun.*, **69**, 4195–4201.
5. Gandhi,N.R., Moll,A., Sturm,A.W., Pawinski,R., Govender,T., Lalloo,U., Zeller,K., Andrews,J. and Friedland,G. (2006) Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet*, **368**, 1575–1580.
6. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. III *et al.* (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.
7. Catanho,M., Mascarenhas,D., Degrave,W. and Miranda,A.B. (2006) GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. *Genet. Mol. Res.*, **5**, 115–126.
8. Vishnoi,A., Srivastava,A., Roy,R. and Bhattacharya,A. (2008) MGDD: Mycobacterium tuberculosis genome divergence database. *BMC Genomics*, **9**, 373–376.
9. Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F., Zachariah,Z.K. *et al.* (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.*, **35**, D766–D770.
10. Galagan,J.E., Calvo,S.E., Borkovich,K.A., Selker,E.U., Read,N.D., Jaffe,D., FitzHugh,W., Ma,L.J., Smirnov,S., Purcell,S. *et al.* (2003) The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, **422**, 859–868.
11. Galagan,J.E., Nusbaum,C., Roy,A., Endrizzi,M.G., Macdonald,P., FitzHugh,W., Calvo,S., Engels,R., Smirnov,S., Atnoor,D. *et al.* (2002) The genome of M. acetivorans reveals extensive metabolic and physiological diversity. *Genome Res.*, **12**, 532–542.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
14. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
15. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
16. Squires,B., Macken,C., Garcia-Sastre,A., Godbole,S., Noronha,J., Hunt,V., Chang,R., Larsen,C.N., Klem,E., Biersack,K. *et al.* (2008) BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Res.*, **36**, D497–D503.
17. Cole,S.T. (1999) Learning from the genome sequence of Mycobacterium tuberculosis H37Rv. *FEBS Lett.*, **452**, 7–10.
18. Terwilliger,T.C., Park,M.S., Waldo,G.S., Berendzen,J., Hung,L.W., Kim,C.Y., Smith,C.V., Sacchettini,J.C., Bellinzoni,M., Bossi,R. *et al.* (2003) The TB structural genomics consortium: a resource for Mycobacterium tuberculosis biology. *Tuberculosis*, **83**, 223–249.
19. Sherman,D.R., Voskuil,M., Schnappinger,D., Liao,R., Harrell,M.I. and Schoolnik,G.K. (2001) Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha-crystallin. *Proc. Natl Acad. Sci. USA*, **98**, 7534–7539.
20. Rees,C.A., Demeter,J., Matese,J.C., Botstein,D. and Sherlock,G. (2004) GeneXplorer: an interactive web application for microarray data visualization and analysis. *BMC Bioinformatics*, **5**, 141.
21. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
22. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. and Mesirov,J.P. (2005) Gene set enrichment

analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
23. Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.
24. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.